OVERVIEW

- STATISTICS

  PANIK ...THE THEORY AND METHODS OF COLLECTING, ORGANIZING, PRESENTING, ANALYZING, AND INTERPRETING DATA SETS SO AS TO DETERMINE THEIR ESSENTIAL CHARACTERISTICS...

  HYPERSTAT ...IN THE BROADEST SENSE, "STATISTICS" REFERS TO A RANGE OF TECHNIQUES AND PROCEDURES FOR ANALYZING DATA, INTERPRETING DATA, DISPLAYING DATA, AND MAKING DECISIONS BASED ON DATA...

- DESCRIPTIVE - SUMMARIZE ESSENTIAL FEATURES OF DATA (CENTRAL TENDENCY, VARIABILITY, DISTRIBUTION)

  INFERENTIAL - ESTIMATES, PREDICTIONS, PREDICTIONS, FORECASTS, GENERALIZATIONS

  PANIK ...INDUCTIVE STATISTICS, INFERRING SOMETHING ABOUT THE WHOLE FROM THE EXAMINATION OF ONLY ONE PART...

- POPULATION

  DANIEL ...LARGEST COLLECTION OF ENTITIES FOR WHICH WE HAVE AN INTEREST AT A PARTICULAR TIME...

  HYPERSTAT ...ENTIRE SET OF OBJECTS, OBSERVATIONS, OR SCORES THAT HAVE SOMETHING IN COMMON...EXAMPLE, A POPULATION MIGHT BE DEFINED AS ALL MALES BETWEEN THE AGES OF 15 AND 18...THE DISTRIBUTION OF A POPULATION CAN BE DESCRIBED BY SEVERAL PARAMETERS SUCH AS THE MEAN AND STANDARD DEVIATION... ESTIMATES OF THESE PARAMETERS TAKEN FROM A SAMPLE ARE CALLED STATISTICS...

- CENSUS
  ...COUNT EVERY ENTITY IN THE POPULATION...

- SAMPLE
  ...PART OF A POPULATION...

  HYPERSTAT...SUBSET OF A POPULATION...SINCE IT IS USUALLY IMPRACTICAL TO TEST EVERY MEMBER OF A POPULATION, A SAMPLE FROM THE POPULATION IS TYPICALLY THE BEST APPROACH AVAILABLE...

- PARAMETER (POPULATION)

  HYPERSTAT...NUMERICAL QUANTITY MEASURING SOME ASPECT OF A POPULATION OF SCORES...EXAMPLE, THE MEAN IS A MEASURE OF CENTRAL TENDENCY...PARAMETERS ARE RARELY KNOWN AND ARE USUALLY ESTIMATED BY STATISTICS COMPUTED IN SAMPLES...

- STATISTIC (SAMPLE)

  HYPERSTAT ..."STATISTIC" IS DEFINED AS A NUMERICAL QUANTITY (SUCH AS THE MEAN) CALCULATED IN A SAMPLE. SUCH STATISTICS ARE USED TO ESTIMATE PARAMETERS...

  HYPERSTAT ..."STATISTICS" SOMETIMES REFERS TO CALCULATED QUANTITIES REGARDLESS OF WHETHER OR NOT THEY ARE FROM A SAMPLE...EXAMPLES, BATTING AVERAGE, "GOVERNMENT STATISTICS"

- QUANTITATIVE (TEMPERATURE)
  QUALITATIVE (GENDER)

  THE QUALITATIVE-QUANTITATIVE DEBATE

  HYPERSTAT ...QUALITATIVE VARIABLES ARE SOMETIMES CALLED "CATEGORICAL VARIABLES"...QUANTITATIVE VARIABLES ARE MEASURED ON AN ORDINAL, INTERVAL, OR RATIO SCALE...QUALITATIVE VARIABLES ARE MEASURED ON A NOMINAL SCALE...

- DISCRETE (NUMBER OF HOSPITAL ADMITS IN A DAY)
  CONTINUOUS (HEIGHT)

  HYPERSTAT...SOME VARIABLES (SUCH AS REACTION TIME) ARE MEASURED ON A CONTINUOUS SCALE...THERE IS AN INFINITE NUMBER OF POSSIBLE VALUES THESE VARIABLES CAN TAKE ON...OTHER VARIABLES CAN ONLY TAKE ON A LIMITED NUMBER OF VALUES AND SUCH VARIABLES ARE CALLED "DISCRETE" VARIABLES...

LEVELS OF MEASUREMENT

- NOMINAL (GENDER, ETHNICITY, RACE)

  HYPERSTAT…

  … NOMINAL MEASUREMENT CONSISTS OF ASSIGNING ITEMS TO GROUPS OR CATEGORIES
  … NO QUANTITATIVE INFORMATION IS CONVEYED AND NO ORDERING OF THE ITEMS IS IMPLIED
  … NOMINAL SCALES ARE QUALITATIVE RATHER THAN QUANTITATIVE
  … FREQUENCY DISTRIBUTIONS ARE USUALLY USED TO ANALYZE DATA MEASURED ON A NOMINAL SCALE
  … MAIN STATISTIC COMPUTED IS THE MODE
  … REFERRED TO AS CATEGORICAL OR QUALITATIVE VARIABLES

- ORDINAL (ATTITUDE SCALE: 0, 1, 2, 3 ,4 5)

  HYPERSTAT...
  ... ORDERED IN THE SENSE THAT HIGHER NUMBERS REPRESENT HIGHER VALUES
  ... INTERVALS BETWEEN THE NUMBERS ARE NOT NECESSARILY EQUAL
  ... NO "TRUE" ZERO POINT FOR ORDINAL SCALES SINCE THE ZERO POINT IS CHOSEN ARBITRARILY

- INTERVAL (TEMPERATURE)

  HYPERSTAT...
  ... ONE UNIT ON THE SCALE REPRESENTS THE SAME MAGNITUDE ON THE TRAIT OR CHARACTERISTIC BEING MEASURED ACROSS THE WHOLE RANGE OF THE SCALE
  ... NO "TRUE" ZERO POINT

- RATIO (HEIGHT)

  HYPERSTAT...
  ... RATIO SCALES ARE LIKE INTERVAL SCALES EXCEPT THEY HAVE TRUE ZERO POINTS

RANDOM SAMPLING

- RANDOM SAMPLE
  ...EACH ***MEMBER OF A POPULATION*** HAS THE SAME CHANCE OF BEING SELECTED...

- SIMPLE RANDOM SAMPLE
  ...EACH SAMPLE OF SIZE N IS SELECTED SUCH THAT EACH SUCH ***SAMPLE OF SIZE N*** HAS THE SAME CHANCE OF BEING SELECTED...

OTHER SAMPLE TYPES

- SYSTEMATIC
  … Nth ENTITY

- STRATIFIED
  … RANDOM ENTITIES WITHIN STRATA (GENDER, AGE GROUP, COUNTY)

- CLUSTER
  … ALL ENTITIES WITHIN RANDOMLY SAMPLED CLUSTERS (COUNTY, CENSUS TRACT, ZIP CODE)

- CONVENIENCE
  … WHATEVER

CRITICAL THINKING CHAPTER 1

AVERAGE AGE AT DEATH FOR VARIOUS OCCUPATIONS

STUDENTS...MEAN = 20.7

MOST DANGEROUS OCCUPATION = STUDENT

???

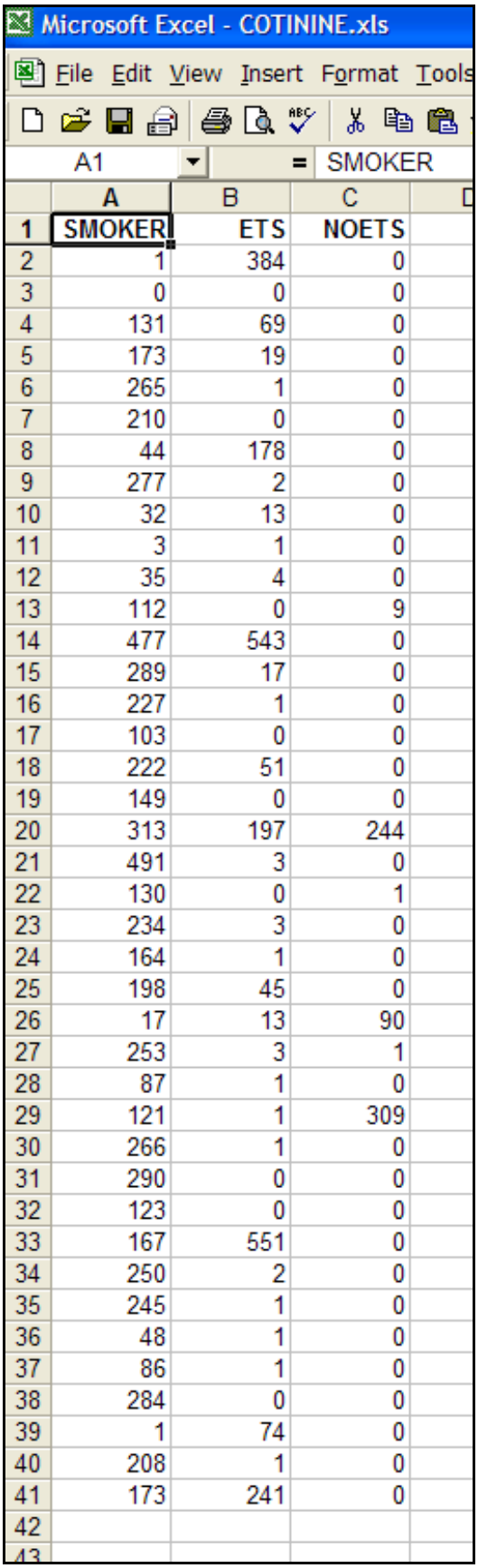
DESCRIBING & COMPARING DATA

- TRIOLA...ONE DATA SET TO ILLUSTRATE CONCEPTS IN CHAPTER

  SERUM COTININE LEVELS IN BLOOD (PRODUCED BY NICOTINE)...INDICATOR OF EXPOSURE TO CIGARETTE SMOKE

- FIVE (WELL, THREE + TWO) IMPORTANT CHARACTERISTICS OF DATA...

  CENTER VARIATION DISTRIBUTION OUTLIERS TIME



	A	B	C
1	SMOKER	ETS	NOETS
2	1	384	0
3	0	0	0
4	131	69	0
5	173	19	0
6	265	1	0
7	210	0	0
8	44	178	0
9	277	2	0
10	32	13	0
11	3	1	0
12	35	4	0
13	112	0	9
14	477	543	0
15	289	17	0
16	227	1	0
17	103	0	0
18	222	51	0
19	149	0	0
20	313	197	244
21	491	3	0
22	130	0	1
23	234	3	0
24	164	1	0
25	198	45	0
26	17	13	90
27	253	3	1
28	87	1	0
29	121	1	309
30	266	1	0
31	290	0	0
32	123	0	0
33	167	551	0
34	250	2	0
35	245	1	0
36	48	1	0
37	86	1	0
38	284	0	0
39	1	74	0
40	208	1	0
41	173	241	0
42			
43			

FROM TRIOLA, CHAPTER 2

COTININE: METABOLITE OF NICOTINE - WHEN NICOTINE IS ABSORBED BY THE BODY, COTININE IS PRODUCED

SMOKER: REPORTED TOBACCO USER

ETS: NON-SMOKERS EXPOSED TO 2ND HAND SMOKE

NOETS: NON-SMOKERS, NO EXPOSURE TO 2ND HAND SMOKE

- FREQUENCY DISTRIBUTIONS

  DATA VALUES GROUPED INTO INTERVALS

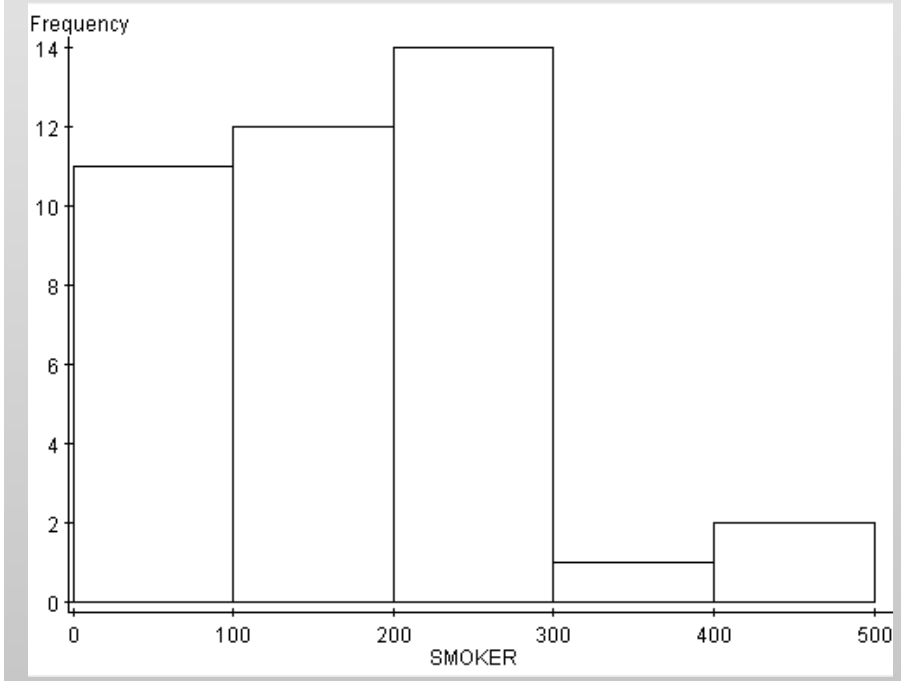
  CHOICE OF INTERVALS AFFECTS HOW YOU (OTHERS) SEE (INTERPRET) YOUR DATA...NO ONE RIGHT ANSWER AS TO WHAT INTERVALS ARE APPROPRIATE

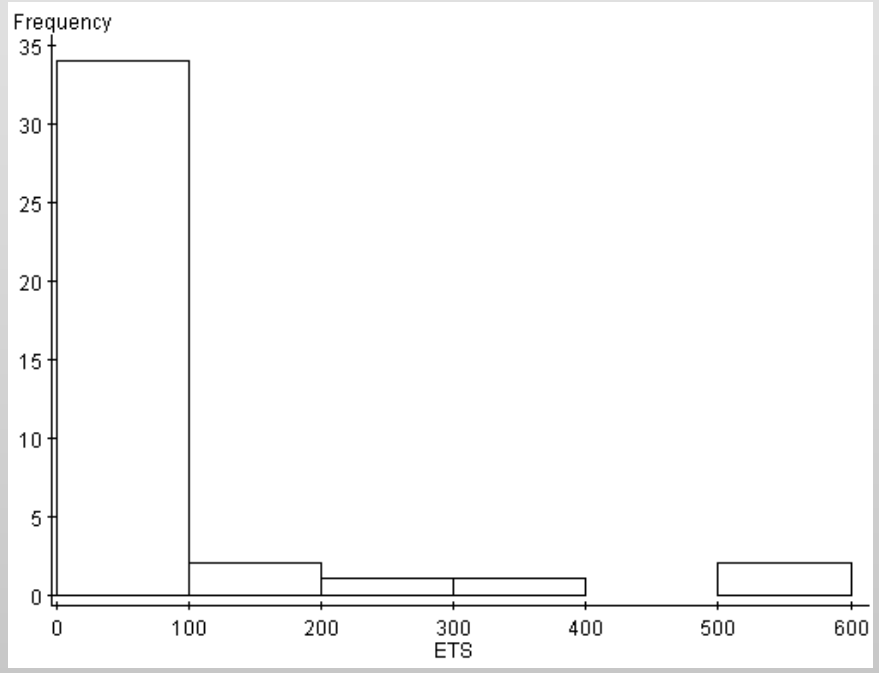
  USER-DEFINED QUARTILES NATURAL BREAKS
  STURGES RULE SOFTWARE-DEFINED

  TABLES OR GRAPHICS

Using Statcrunch

Histograms



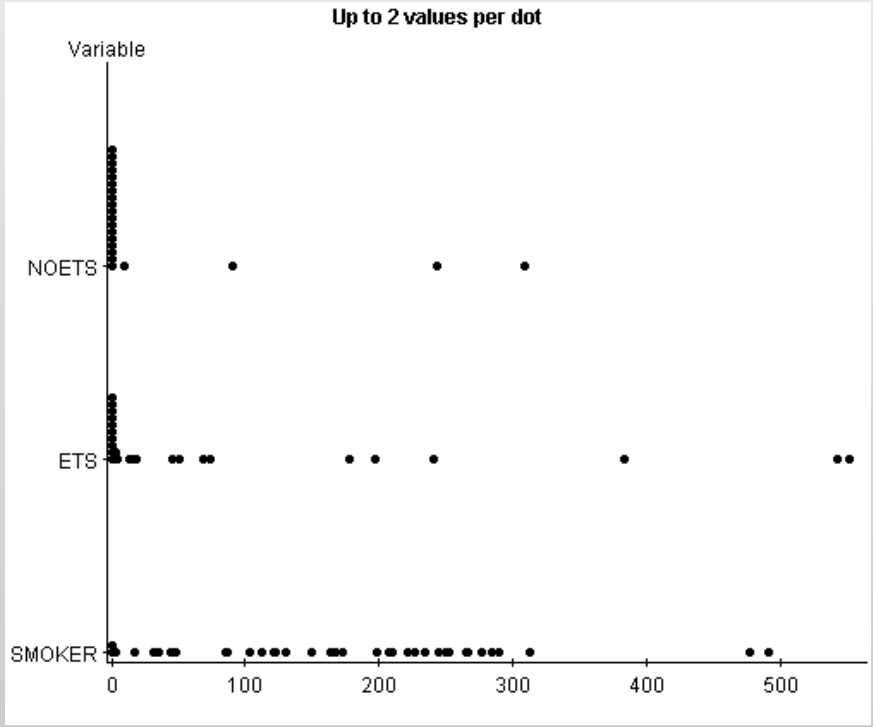


Using Statcrunch

Stem-and-Leaf

Variable: SMOKER
0 : 00002344
0 : 599
1 : 012233
1 : 56777
2 : 011233
2 : 555778899
3 : 1
3 :
4 :
4 : 89

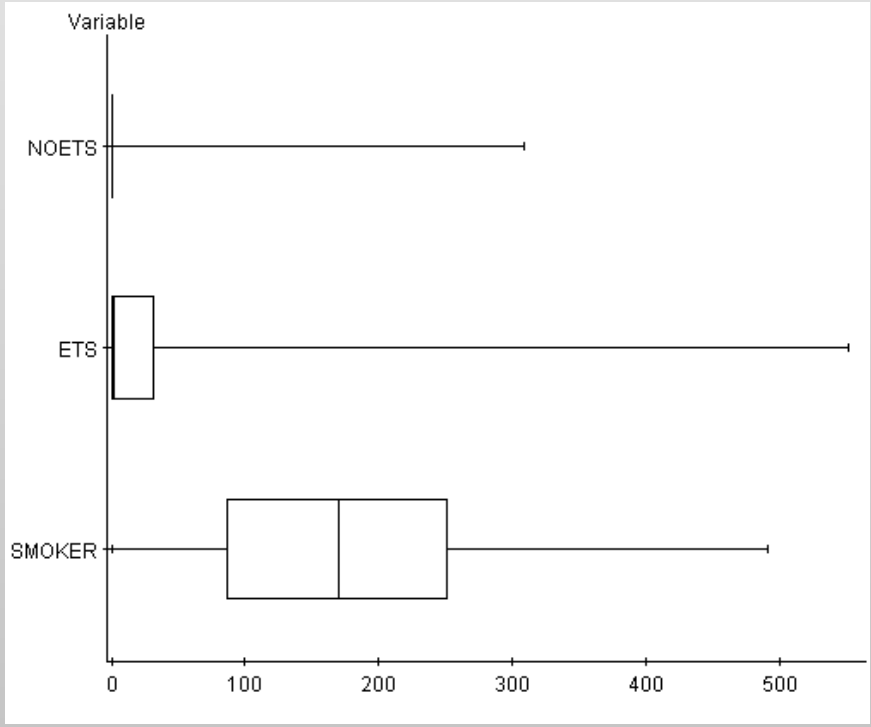
Dot plot



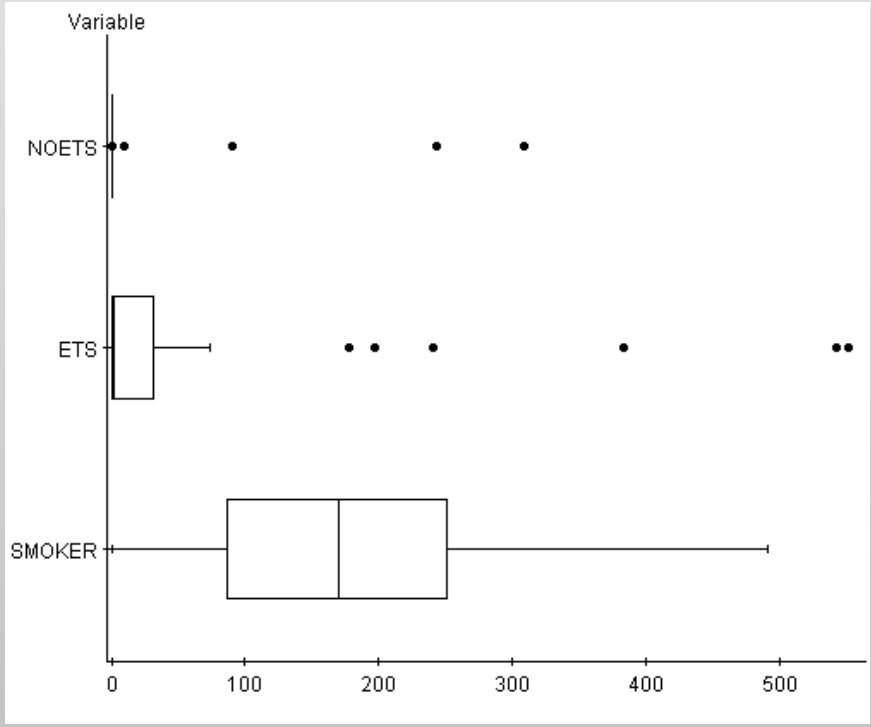
Using Statcrunch

Boxplots

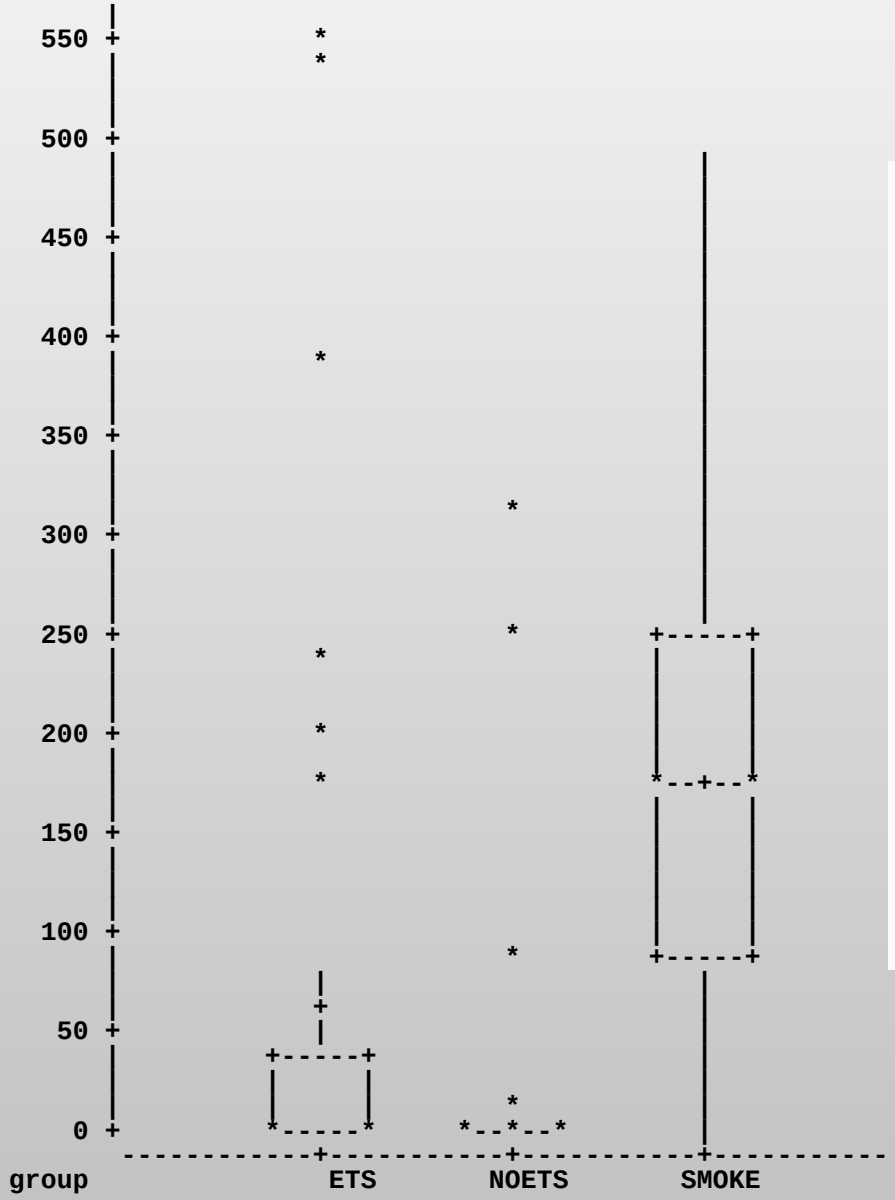
No Fences

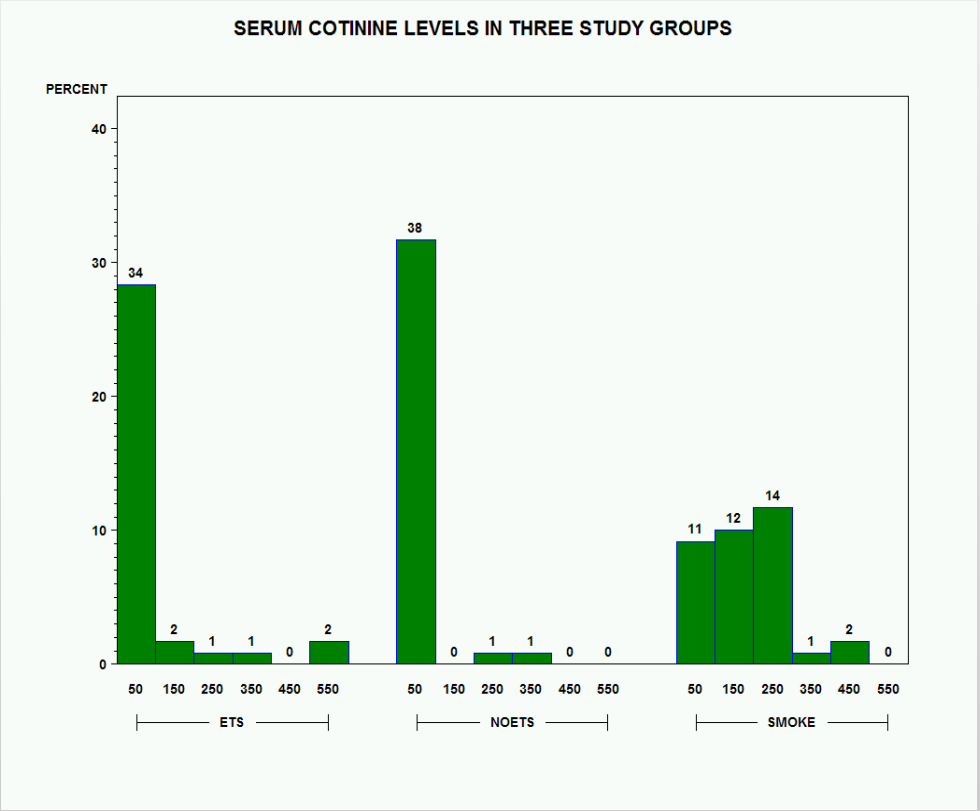


With Fences



Using SAS


550
500
450
400
350
300
250
200
150
100
50
0
group
ETS
NOETS
SMOKE


SERUM COTININE LEVELS IN THREE STUDY GROUPS
PERCENT
40
30
20
10
0
34
2
1
1
0
2
38
0
1
1
0
0
11
12
14
1
2
0
50 150 250 350 450 550
ETS
NOETS
SMOKE

- OTHER GRAPHICS

  SCATTER PLOTS (2 VARIABLES)

  TIME SERIES (TREND LINES)

  OGIVE (CUMULATIVE FREQUENCY)

  MISCELLANEOUS (USER CREATIVITY !!!)

- MEASURES OF CENTER

MEAN — SUM OF ALL VALUES DIVIDED BY THE NUMBER OF VALUES (VERY SENSITIVE TO EXTREME VALUES)

MEDIAN — MIDPOINT - NOTHING TO DO WITH THE VALUES THEMSELVES (NOT SENSITIVE TO EXTREME VALUES)

MODE — MOST FREQUENTLY OCCURRING VALUE (POSSIBILITY OF MULTIPLE MODES, BIMODAL NOT UNCOMMON)

RELATIONSHIP OF MEAN, MEDIAN, MODE (SYMMETRIC AND SKEWED DISTRIBUTIONS)

- MEASURES OF VARIATION

RANGE	DIFFERENCE BETWEEN MINIMUM AND MAXIMUM VALUES
STANDARD DEVIATION	MEASURE OF VARIATION OF VALUES AROUND THE MEAN
VARIANCE	AVERAGE SQUARED DEVIATION OF A VALUE FROM THE MEAN (SQUARE OF STANDARD DEVIATION) - DENOMINATOR, N VERSUS N-1 AND BIAS
COEFFICIENT OF VARIATION	STANDARD DEVIATION EXPRESSED AS A PERCENTAGE OF THE MEAN

- STANDARD DEVIATION AND 'TYPICAL VALUES'

  EMPIRICAL RULE - IN A BELL-SHAPED DISTRIBUTION...

  68% OF VALUES FALL WITHIN 1 STANDARD DEVIATION OF THE MEAN...95% OF VALUES FALL WITHIN 2 STANDARD DEVIATIONS OF THE MEAN...99.7% OF VALUES FALL WITHIN 3 STANDARD DEVIATIONS OF THE MEAN

  ACCEPTED CONCEPT OF 'TYPICAL VALUES' --- WITHIN 2 STANDARD DEVIATIONS OF THE MEAN

- CHEBYSHEV'S THEOREM

  REGARDLESS OF THE SHAPE OF A DISTRIBUTION, THE PROPORTION OF DATA LYING WITHIN K STANDARD DEVIATIONS OF THE MEAN IS *AT LEAST*...$1 - 1/K^2$

  AT LEAST 3/4 (75%) OF ALL VALUES LIE WITHIN 2 STANDARD DEVIATIONS OF THE MEAN

  AT LEAST 8/9 (89%) OF ALL VALUES LIE WITHIN 3 STANDARD DEVIATIONS OF THE MEAN

- RATIONALE FOR STANDARD DEVIATION (AND VARIANCE)

- RELATIVE STANDING

  Z-SCORES STANDARDIZED SCORES USING MEAN AND STANDARD DEVIATION
  (UNUSUAL VALUES - JORDAN, LOBO, BOGUES - OUTSIDE OF +/- Z=2)

  QUARTILES Q1, Q2, Q3
  MEDIAN, MEDIANS OF UPPER AND LOWER HALF
  INTERQUARTILE RANGE (Q1 AND Q3)

- EDA (EXPLORATORY DATA ANALYSIS - JOHN TUKEY, 1977)

  EDA IS DETECTIVE WORK (NUMERICAL, COUNTING, GRAPHICAL)

  EDA CAN NEVER BE THE WHOLE STORY, BUT NOTHING ELSE CAN SERVE AS THE FOUNDATION STONE, AS THE 1ST STEP

  SEARCH FOR PATTERNS, GROUPS, UNEXPECTEDLY POPULAR VALUES, UNUSUAL VALUES, OUTLIERS

  TRIOLA 5-NUMBER SUMMARY (MINIMUM, Q1, MEDIAN, Q3, MAXIMUM)

  STEM-AND-LEAF PLOTS, BOX PLOTS

JOHN W. TUKEY
Princeton University and
Bell Telephone Laboratories

Exploratory Data Analysis

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts • Menlo Park, California
London • Amsterdam • Don Mills, Ontario • Sydney

Value	Break	Value
−240	617	−160
−236	633	−156
−232	649	−152
−228	666	−148
−224	685	−144
−220	704	−140
−216	725	−136
−212	746	−132
−208	769	−128
−204	793	−124
−200	820	−120
−196	840	−118
−192	855	−116
−188	870	−114
−184	885	−112
−180	901	−110
−176	917	−108
−172	935	−106
−168	952	−104
−164	971	−102
	990	

scale values

We ought to put as many scale values on the graph paper preliminary as will help us make the plot easily. On the tracing paper final, however, we ought not show more than three or four numbers along a scale. More clutters up the picture and distracts the eye from what it ought to see. (Scales for dates are sometimes an exception. It can matter whether an appearance came in 1929 or 1928, in 1776 or 1775.)

People are used to scales on the left and below. So be it--for the picture, perhaps. When one is plotting the points, however, it is much more convenient to put the horizontal scale ABOVE the plot, where you do not have to move your hand to see it. (It would be rational to plot from detailed scales above and left, and to produce a final picture with a few scale points shown below and right; but such rationality is usually not worth the possibility of occasional confusion.)

plotting without graph paper

We almost always want to look at numbers. We do not always have graph paper at hand. **There is no excuse for failing to plot and look.**

We usually have ruled paper at hand. For emergency graph paper, take out one sheet of ruled paper, turn it on its side, and place it beneath another sheet of ruled paper. If these two sheets have a light-colored backing--often provided by the rest of the pad or notebook--the vertical lines on the lower sheet are almost certain to show through well enough, combining with the horizontal lines on the top sheet to form a grid on which plotting is reasonably easy. (The first step in this sort of plotting is to mark--by ticks or unobtrusive dots--enough information on the top sheet to make it easy to get the lower sheet back to its original position after it slips.)

With this technique, one can make useful, if not decorative, plots almost anywhere.

review questions

What is a box-and-whisker plot? What do its parts show forth? What rules does it obey about showing values individually? About identifying values? What must we separate in our minds about plotting? What are the essentials of convenient, effective plotting? How can we, in an emergency, plot without graph paper?

2D. Fences, and outside values

Hinges are for our convenience. They can--and will--serve various purposes for us. Their role in 5-number summaries is only the beginning.

When we look at some batches of values, we see certain values as apparently straying out far beyond the others. In other batches straying is not so obvious, but our suspicions are alerted. It is convenient to have a rule of

thumb that picks out certain values as "outside" or "far out". To do this, we set up appropriate "fences" and use "outside" and "far out" accordingly.

A useful rule of thumb runs as follows:

- ◇"H-spread" = difference between values of hinges.
- ◇"step" = 1.5 times H-spread.
- ◇"inner fences" are 1 step outside hinges.
- ◇"outer fences" are 2 steps outside hinges (and thus 1 step outside of inner fences).
- ◇the value at each end closest to, but still inside, the inner fence is "adjacent".
- ◇values between an inner fence and its neighboring outer fence are "outside".
- ◇values beyond outer fences are "far out".

Exhibit **7** shows some examples. Notice the practice of writing the value of the H-spread just outside the box (under a horizontal "eave"), and then putting the value of the step in a "penthouse" on the second part of the display, which contains the fences and is written after the main letter display, either below it or to its right. Panels C to F show a convenient standard form, combining summary scheme, fences, and identification of outside values. We can call it a

fenced letter display

exhibit **7** of chapter 2: various examples

Examples of calculation of fences and identification of adjacent, outside, and far outside values (based on exhibits 3 and 4)

A) For panel A of exhibit 3

17 Chev. prices

M9	895		
H5	795	1499	704
1	150	1895	

	1056		
f	−261	2555	
	xxx	xxx	out
F	−1317	3611	
	xxx	xxx	far

adj: 150, 1895

Note that, here:

704 = 1499 − 795
1056 = (1.5)(704)
−261 = 795 − 1056
2555 = 1499 + 1056
−1317 = −261 − 1056
3611 = 2555 + 1056

→

Origin of the term [edit]

This has been said to derive from the belief that English law allowed a man to beat his wife with a stick so long as it is was no thicker than his thumb. In 1782 Judge Sir Francis Buller is reported as having made this legal ruling. The following year James Gillray published a satirical cartoon attacking Buller and caricaturing him as 'Judge Thumb'. The cartoon shows Buller carrying two bundles of sticks and the caption reads "thumbsticks - for family correction: warranted lawful!"

Caricature condemning Buller: *Judge Thumb - Patent Sticks for Family Correction - Warranted Lawful!*

It seems that Buller was hard done by. He was notoriously harsh in his punishments, but there's no evidence that he ever made the ruling that he is infamous for. Edward Foss, in his authoritative work The Judges of England, 1870, wrote that, despite a searching investigation, "no substantial evidence has been found that he ever expressed so ungallant an opinion".

It's certainly the case that, although British common law once held that it was legal for a man to chastise his wife in moderation (whatever that meant), the 'rule of thumb' has never been the law in England. Despite the phrase being in common use since the 17th century and appearing many thousands of times in print, there are no printed records that associate it with domestic violence until the 1970s. The false stories that assumed the wife-beating law to be true may have been influenced by Gillray's cartoon.

Even if people mistakenly believed that law to exist, there's no reason to connect the legal meaning with the phrase - which has been in circulation since at least 1692, when it appeared in print thus:

Sir W. Hope, Fencing-Master, 1692 - "What he doth, he doth by rule of Thumb, and not by Art."

That makes it clear that the origin refers to one of the numerous ways that thumbs have been used to estimate things - judging the alignment or distance of an object by holding the thumb in one's eye-line, the temperature of brews of beer, measurement using the estimated inch from the joint to the nail, etc. It isn't clear which of these is the precise origin and this joins the whole nine yards as a phrase that probably derives from some indetermined form of measurement.